DOCUMENT RESUME

ED 269 681 CG 019 066

AUTHOR Waln, Ronald F.; Downey, Ronald G.
TITLE Voice Stress Analysis: Use of Telephone Recordings.
PUB DATE [Aug 85]
NOTE 25p.; Paper presented at the Annual Convention of the American Psychological Association (93rd, Los Angeles, CA, August 23-27, 1985).
PUB TYPE Reports - Research/Technical (143) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Employment Interviews; Job Applicants; *Lying; Personnel Selection; *Polygraphs; *Stress Variables; *Tape Recordings; *Telephone Communications Systems
IDENTIFIERS *Voice Stress Analysis

ABSTRACT

The ability to detect lying is an important skill. While the polygraph is the most common mechanical method used for lie detection, other electronic-based methods have also been developed. One such method, the analysis of voice stress patterns, is based on the assumption that lying is a stressful activity which reduces involuntary frequency modulations in the human voice. One variation of voice analysis involves recording interviews and then transmitting the recordings through the telephone to a second location where the voice is re-recorded, charted, and evaluated. Voice stress analyses were performed on 15 tape-recorded pre-employment interviews in both their original form and after they had been transmitted via telephone and re-recorded. Four expert voice stress examiners, blind to the telephone condition, reported less stress in the telephone charts than in the original charts. There was little relationship between the stress rating for the same charts in their original and telephone forms. Reliability estimates were low for both the original and telephone stress ratings. Summing over the stress ratings from individual questions and advanced training on the part of the examiners both appeared to improve the reliability estimates. The continued use of telephone recorded tapes as substitutes for the original tapes is highly questionable. In addition, these results suggest that voice analysis ratings, as they are currently used, do not show sufficient reliability to warrant their continued use as a selection procedure for employment. (NB)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED269681

Voice Stress Analysis:
Use of Telephone Recordings

Ronald F. Waln and Ronald G. Downey

Kansas State University

CG 019066

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

Ronald G. Downey

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

Running head: Voice Stress Analysis

Abstract

Voice stress analyses were performed on tape recorded pre-employment interviews in both their original form and after they had been transmitted via telephone and re-recorded. Expert voice stress examiners, blind to the telephone condition, reported less stress in the telephone charts than in the original charts. There was little relationship between the stress rating for the same charts in their original and telephone forms. Reliability estimates were low for both the original and telephone stress ratings. Summing over the stress ratings from individual questions and advanced training on the part of the examiners both appeared to improve the reliability estimates. The continued use of telephone recorded tapes as substitutes for the original tapes is highly questionable. In addition, these results suggest that voice analysis ratings, as they are currently used, do not show sufficient reliability to warrant their continued use as a selection procedure for employment.

Voice Stress Analysis:
Use of Telephone Recordings

The ability to detect deception and lying is an important and much sought out skill. Lie detection plays a role in civil and criminal court cases (Kleinmuntz & Szucko, 1982 and Lykken, 1984); industrial settings (Bell, 1981;Lykken, 1981; Sackett & Decker, 1979); and government settings (Mervis, 1983 and Saxe, Dougherty, & Cross, 1985). While the polygraph is the most common mechanical method in use for lie detection (Kleinmuntz & Szucko, 1984; Sackett & Decker, 1979), other electronic-based methods have recently emerged. One such system involves the analysis of voice stress patterns, a technique which has been subjected to only limited study by psychologists (Sackett & Decker, 1979).

Voice analysis is based on the assumption that lying is a stressful activity which reduces involuntary frequency modulations in the human voice (Dektor Counterintelligence and Security, Inc.,[Dektor] 1971). Dektor claims that vocal modulations are detected, measured, and displayed by the voice analysis equipment and examiners can be trained to interpret these displays. The use of voice analysis techniques for identification of lying is reported to have several advantages over polygraph procedures, including eliminating the need for direct physical hookups, the ability to use recordings taken without the knowledge of the person, the versatility of being able to conduct the interview in almost any location (restricted only by being able

to use a tape recording), and being able to transmit recordings of the interview over the telephone for evaluation elsewhere (Dektor, 1971). The increasing use of voice stress lie detection methods in government and industry (Bell, 1981), and the potential for abuse of voice stress analysis (Hollien, 1980) indicates the need to investigate the various claims being made concerning this new technique.

While proponents (e.g., Bell, 1981) have claimed that voice stress analysis techniques are at least as good as polygraphs in detecting deception, laboratory evidence is at this time equivocal. Horvath (1978) and Kubis (1973) both reported that voice stress analysis produced approximately chance level identification of lying in mock crime situations, while the polygraph equipment performed well beyond chance levels. Horvath also reported a correlation of .38 between the two voice stress examiners. In a follow up study Horvath (1979) put additional stress on the subjects by only awarding extra credit if the subjects were successful in either being caught or avoiding being caught lying. Once again the "hit rates" for voice stress testing were no better than chance but the correlation between the examiners was higher, $\underline{r}$=.65. Both Bell (1981) and Heisse (1976) have asserted that the lack of positive findings is due to the generally non-risk nature of the experimental setting, and maintain that only in real world situations can voice stress equipment be tested fairly. A similar argument has been used to explain the negative results from laboratory studies of

polygraphs (Lykken, 1979).

Attempts to improve the ecological validity of voice stress research have produced more positive results, but in less controlled situations. Kradz (1971) tape recorded 42 polygraph interviews with suspects or victims of actual crimes. Blind evaluations of the voice analysis charts agreed with polygraph results in all but one case. In two separate voice analysis evaluations the examiners agreed perfectly. Kradz also reported that the final dispositions of the cases were observed and collaborated the results of the polygraph examinations. Heisse (1976) collected 53 voice analysis interviews acquired during actual criminal or pre-employment investigations. The final dispositions of these cases were known, usually through confessions. Two examiners blindly rated the voice stress data using a standardized evaluation method developed by Heisse (1974). Heisse (1976) reported that 97% of the examiners' ratings were correct, and the interrater reliability was .96. He concluded that the use of standardized methods in a non-experimental environment produced these very positive results.

At a more basic level, attempts to demonstrate that voice analysis evaluations can detect stress have produced mixed results. Lynch and Henry (1979) found voice stress evaluators unable to correctly identify responses to taboo versus non-taboo words. However, VanDercar, Greaner, Hibler, Spielberger, and Bloch (1980) found that voice stress evaluation identified changes in state anxiety (State-Trait Anxiety Inventory, Spiel-

berger, Gorsuch, & Lushene, 1970) only when the threat of shock or taboo words was high. They reported an interrater reliability for the four raters of .92. Brenner, Branscomb, and Schwartz (1979) found that voice analysis of individuals varied as a function of the task difficulty associated with mathematic problems they were solving. These results suggest that the voice stress analysis technique may have some validity with high stress stimuli.

While the above review of the research on voice stress analysis suggests some value for the technique, practitioners use the procedures for lie detection in a variety of situations in which our knowledge is sorely lacking. One of the areas where practice may have exceeded our understanding is in the use of the telephone for transmitting voice recordings of interviews. The use of telephone transmissions allows an interview to be conducted and recorded at one site, then the recording can be played through the telephone and re-recorded at another location where it can be charted by the equipment and evaluated by the examiners. Dektor (1971), the manufacturer of the Psychological Stress Evaluator (P.S.E.), claims that their device works as well using telephone recordings as it does with the original recording. There are several reasons for suggesting that this claim may not be correct. The P.S.E. is believed to detect frequency modulations in the 8-14 HZ region while the telephone transmits frequencies in the 300-3300 HZ range. Further, there is the potential for a variety of noise to be introduced into the

recordings by the transmission process. These problems have been previously identified by Hollien (1980) but no attempts to investigate this issue were found. In addition to the telephone transmission concerns, there is a need to provide further work on the reliability and validity of the voice stress analysis approach.

The primary purpose of this study was to determine if telephone recorded tapes of pre-employment interviews and non-telephone recordings of the same interviews were evaluated in a similar way. Reliability will be determined for each method and for different types of questions.

Method

Subjects

Tape recorded interviews were selected from the files of a midwest security consulting company which routinely conducts both in-person and telephone-transmitted pre-employment P.S.E. interviews. The 15 subjects whose records were used had applied for sales positions with the same retail organization. Each of the subjects had undergone an in-person P.S.E. examination and each had been judged deceptive by the original examiner.

Four professional P.S.E. examiners agreed to rate the P.S.E. charts (the tracings of the frequency modulations). Although the qualifications of the raters varied, all four had completed an approved training program in P.S.E. chart interpretation and had field experience in chart interpretation. The raters received the charts in the mail and returned them by mail after making their evaluations.

Equipment

The original interviews of the subjects were recorded using a Uher 4000IC reel to reel tape recorder. These recorded interviews were charted using a P.S.E. model 101. For the telephone charts the following procedures were followed. The original tapes of the interview were transmitted to the security company over the telephone. While the manufacture of the P.S.E. provides accessories for use with the telephone, the security firm uses its own equipment. This equipment consists of a Superscope C-202LP cassette tape recorder which is wired directly into a standard telephone (by-passing the handset) at the origin of the transmission and a Uher 4000IC recorder wired directly into the telephone at the terminal end of the transmission. While telephone transmissions are often done on a long distance line, the telephone tapes were produced using a local line. The Uher 4000IC recording was used to produce the chart.

Procedures

Two sets of P.S.E. charts were evaluated by each of the raters. One set consisted of the "Regular" 15 charts taken directly from the recordings of the interviews, and the second set of the "Telephone" charts produced by the procedures described above. The charts from the two sets were presented in a random order with the restriction that a Regular and Telephone chart from the same person could not be presented one after the other. The raters were informed that the charts were from 30 different individuals who had all responded to the same 23 ques-

tions. The raters were unaware of the nature of the research question. Stress was rated on a 5 point scale: (1) little or no stress indicated; (2) a small, but noticeable amount of stress is present; (3) a moderate amount of stress is present which is indicative of more than "general nervousness"; (4) heavy stress, the question evoked a strong reaction in the subject; and (5) extreme stress, a virtual panic reaction. A final rating scale was included on the form and asked evaluators to rate the degree of overall stress. Two raters declined to use the overall stress scale and it was, therefore, dropped from any of the analyses.

The pre-employment interviews used in this study were conducted using a control question format (Szucko & Kleinmuntz, 1981). In the control question approach the individual is asked "relevant" questions (e.g., Have you ever stolen cash from a previous employer?) and the responses are compared to the response from "irrelevant" questions (e.g., Do you sometimes drive a car?). The "irrelevant" questions are also referred to as "known truth" questions; used to establish an assumed baseline of honest responding. Other control questions are intended to produce stress responses and include the "known lie" and the "outside issues". Comparisons between the charts from various types of questions presumably allow judgments concerning the truthfulness of the responses.

Analysis

The initial analysis was a multivariate ANOVA with the questions used as the dependent variables (this was suggested by

Saal, Downey, and Lahey, 1980). Type of chart (Regular and Telephone), raters, and ratees were the independent variables with 2, 4, and 15 levels respectively. The MANOVA allowed test for determining mean differences between the chart types, a test of the significance of intraclass correlation (Ratee effect), and a test of the degree to which the raters produced different, relatively higher or lower, levels of stress rating. If either chart type and/or an interaction between chart type and another variable was significant, the assessment of reliability from the MANOVA would not be meaningful and a secondary set of Ratee by Rater ANOVAs, one for each question, would be conducted within chart type. These ANOVAs would allow for estimating the intraclass reliabilities within chart type for each question using the appropriate estimate of reliability for a single rater (Shrout & Fleiss, 1980; Model 3,1). If it is assumed that the responses to single questions are all measuring stress and that a summation over the items would be a more reliable measure of this stress, a new score could be computed for each rater. This score was produced by adding the stress ratings for each relevant item together for a rater for each of the 30 charts. Coefficient Alphas were also computed for each rater on each chart type using raters as the test and the 14 relevant questions as the items. Pearson product moment correlations were then computed between the summed ratings (for each rater) for the fifteen ratees and the two chart conditions. The resultant correlation matrix provides a multi-method-multi-rater look at the ratings (Lawler, 1967 and

Campbell & Fiske, 1959). The correlations between raters using the same chart type estimate the interrater reliability of the summed ratings using a particular type of chart. The correlation across chart type for the same rater shows the degree of method convergence. The cross method and rater correlations indicate the degree of convergence over both charts and raters.

Results

All three main effects (chart type, raters, and ratees) for the multivariate analysis of variance were significant and the chart type by ratee interaction effect was also significant. Table 1 gives the multivariate results and the univariate F-Tests for each question. Eleven of the 23 univariate tests were significant for chart type, 14 for the raters, 22 for ratees, and 9 for the chart type by ratee interaction. As a general rule the questions from the telephone charts were rated as showing less stress and this was true for all the questions where the difference was significant. Raters demonstrated a moderate level of reliability (intraclass correlations were computed but are not shown) in the rank ordering of the charts for each question (when averaged over chart type). Fourteen Rater univariate main effects and the multivariate main effect were significant. Raters differed in their ratings of stress over all ratees and chart types. Given the mean differences between chart types and the significant chart type by ratee interactions, it was necessary to conduct separate rater by ratee analyses for each question to make a meaningful assessment of reliability within chart type.

Insert Table 1 about here

Table two summarizes the results of the rater by ratee univariate analyses for each chart type and each question. Questions were organized in Table 2 by question type; relevant, known lie and known truth. For the Regular charts, 17 (out of the 23) questions had significant ratee effects. All of the intraclass correlations were less than .51 and the majority yielded values less than .4. For the Telephone charts, 19 ratee main effects were significant. All of the intraclass correlations were found to be less than .64 and the majority were less than .4. These results indicated that while there was a significant level of interrater reliability, the reliability estimates were quite low. When the ratings (summed over raters) for Regular charts were correlated with the summed ratings for Telephone charts for each question, only 4 out of the 23 resultant correlations were found to be significant (see Table 2).

Insert Table 2 about here

Table 2 demonstrates one other important finding. The differences between the Regular and Telephone charts were more prevalent for the relevant items than they were for the known truth and known lie items; 9 out of 14 relevant items, 1 out of 6 known truth items, and 1 out of 3 known lie items were signifi-

cant. In all the cases where a significant difference occurred between the Regular and Telephone charts, the Regular charts were rated as showing higher stress.

As a final method of determining what was happening, the stress ratings for the 14 Relevant questions were summed over the questions for each rater. This resulted in each rater having a stress score for each individual on each chart type. These 8 different kinds of summated ratings were correlated over 15 ratees (see Table 3). Table 3 also shows the reliability (coefficient alpha) for each rater over the relevant questions and for each chart type. The reliability estimates suggest that, if a rater had rated an individual as high (or low) on a relevant question, then they tended to rate them high (or low) on all the other relevant questions. The circled values in Table 3 represent cross method convergence, the relationship between the ratings for a single rater on one chart with his ratings on the other chart. Raters 1 and 3 both had significant correlations between their rating on one type of chart with their rating on the other type. Only rater 1's correlation (.71) would be considered an acceptable level of convergence. When the correlations among the four raters were examined, both within a chart type and over types of charts, only the ratings from raters 1 and 3 showed consistent significant correlations; Raters 1 and 3 correlated $\underline{r}$=.62 within the Regular condition and .86 in the Telephone condition. Rater 1's ratings in the Regular condition were significantly related to rater 2's ratings in the Telephone

condition (.56) and visa versa (.57). Raters 2 and 4 showed some convergence with rater 3 in the Regular and Telephone conditions respectively but there was no other convergence.

Insert Table 3 about here

Discussion

The research questions asked in this study can, within a limited context, be answered. There is evidence for only a limited degree of interrater reliability for questions from either Regular or Telephone charts. Further, the data demonstrated that the use of Telephone charts led to lower stress ratings and the telephone ratings were not correlated with the ratings from a Regular chart. These findings would argue strongly for discontinuing the use of telephone reproduced tapes as a substitute for regular charts since both mean differences in charts and a lack of convergence was found. Given that the charts used in this study were from actual employment interviews, they are not subject to concerns about their ecological validity (Bell, 1981 and Heisse, 1976).

The analysis of the reliability of single questions from both types of charts was less than encouraging. The intraclass correlations, while generally significant, were low and averaged less than .40. Little ,if any, difference in reliability occurred over the type of question. These estimates are consistent with the .38 value reported by Horvath (1978). Basing important

personnel decisions upon rating where there is so little consensus between two different raters is, to say the least, questionable.

The additional analyses, done with only the Relevant items, offered some hope for improving the reliability of voice analysis ratings. Using the traditional method of summing over items responses (Edwards, 1957), it was apparent that each of the raters were acting in a very consistent fashion for a ratee over items. Further, for raters 1 and 3, the ratings were consistent over the chart conditions as well as within the chart condition. Raters 1 and 3 were the most experienced and had the best training of the 4 raters. These significant correlations between the Regular and Telephone charts (summated ratings) for raters 1 and 3 support the view that some consistent factor in the charts was being observed. However the mean differences for the summated ratings (observed in all four raters) between the Regular and Telephone charts would lead to lower levels of stress (lying) being attributed to the same individual, depending on the source of the charts. Under the best of conditions (assuming the raters are well trained and they are in fact rating stress which is related to lying)) some adjustment would be required in the mean stress levels for charts from telephone tapes. Also, this conclusion would only hold under the situation where stress is evaluated on a series of questions and than added (averaged) over these questions, a procedure which is NOT currently being used.

The results of this study do not provide any direct evidence as to what raters are evaluating. There is, however, some indirect evidence to indicate that the information being evaluated is affected by the content of the question. The mean differences between the Regular and Telephone charts were more often significant for the Relevant items (64%) versus the other items (22%). This suggests that whatever the evaluators were rating, when they were dealing with relevant stress items their ratings were affected by the chart type.If evaluators were rating an individual characteristic unrelated to stress, e.g., voice quality, it would be expected that similar mean differences would be found in both the relevant and the non-relevant questions.

There is little evidence in this study to support the value of voice analysis, as it is currently being used, as a technique for identification of lying, a view shared by Sackett and Decker (1979, p501). The average reliability (interrater) of single raters on single questions was too low to justify the continued use of stress ratings for individual selection purposes. Further, the continued use of tapes transmitted by telephone as a substitute for regular charts, has to be severely questioned given general lack of correlation with regular charts and the lower mean scores given to the Telephone charts. The results obtained from using summated rating over questions offers some future possibilities and offers a systematic alternative to the current practice of asking evaluators to provide a new summary judgment concerning the overall level of stress. If the current methodolo-

gy for evaluating the charts is modified, it results in improvements in the reliability estimates. The individual differences between raters in the interrater reliability estimates needs further investigation and if training is as important as these results suggest, then it is imperative that more time, effort, and resources go into the development and evaluation of training programs.

The use of voice stress analysis techniques and the public's belief in its value as a method for detecting deception has grown over the last few years. As pointed out by Kleinmuntz and Szucko (1984) in their discussion of polygraphic evidence, positive beliefs by the public and supportive pronouncements by proponents/users tend to overwhelm any scientific evidence to the contrary. Since voice analysis results are being used to determine the employability of individuals, and since work is, for most individuals, a major factor in their physical, social, and personal wellbeing, it is mandatory to insure that the reliability and validity of a selection technique is sufficient to warrant its continued use. If voice stress analysis is (as is suggested here) found wanting in reliability, then its use should be discontinued until evidence can be supplied as to its value.

References

Bell, A. D. (1981). The PSE: A decade of controversy. Security Management, 25, 63-73.

Brenner, M., Branscomb, H., & Schwartz, G. (1979). Psychological Stress Evaluator: Two tests of a vocal measure. Psychophysiology, 16, 351-357.

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. Psychological Bulletin, 56, 81-105.

Dektor Counterintelligence and Security, Inc. (1971). Psychological Stress Evaluator. (Available from Dektor CI/S, Inc., 400 Mall Boulevard, Suite M, Savanna, GA, 31406)

Edwards, A. L. (1957). Techniques of attitude scale construction. New York: Appleton-Century Crofts.

Heisse, J. (1976). Audio stress analysis: A validation and reliability study of the Psychological Stress Evaluator. Proceedings of the 1976 Carnahan Conference on Crime Countermeasures, 5-17.

Hollien, H. (1980). Vocal indicators of psychological stress. Annals New York Academy of Sciences, 347, 60-72.

Horvath, F. (1978). An experimental comparison of the Psychological Stress Evaluator and the galvanic skin response in detection of deception. Journal of Applied Psychology, 63, 338-344.

Horvath, F. (1979). The effects of different motivational instructions on detection of deception with the Psychological Stress Evaluator and the galvanic skin response. Journal of Applied Psychology, 64, 323-330.

Kleinmuntz, B., & Szucko, J. J. (1982). On the fallibility of lie detection. Law and Society Review, 17, 85-104.

Kleinmuntz, B., & Szucko, J. J. (1984). Lie detection in ancient and modern times: A call for contemporary scientific study. American Psychologist, 39, 766-776.

Kradz, M. (1971). Psychological Stress Evaluator: A study. Unpublished manuscript. (Available from Dektor CI/S, Inc., 400 Mall Boulevard, Suite M, Savanna, GA, 31406)

Kubis, J. F. (1973). Comparison of voice analysis and polygraph as lie detection procedures (Contract No. DAAD05-72-C-0217). Aberdeen Proving Grounds, MD: U. S. Army Land Warfare Laboratory.

Lawler, E. E. III (1967). The multi-trait-multi-rater approach to measuring managerial job performance. Journal of Applied Psychology, 51, 369-381.

Lykken, D. T. (1979). The detection of deception. Psychological Bulletin, 86, 47-53.

Lykken, D. T. (1981). A tremor in the blood: Uses and abuses of the lie detector. New York: McGraw Hill.

Lykken, D. T. (1984). Trial by polygraph. Behavioral Science and the Law, 2, 75-92.

Lynch, B. E., & Henry, D. R. (1979). A validity study of the Psychological Stress Evaluator. Canadian Journal of Behavioral Science, 11, 89-94.

Mervis, J. (1983, December). Report questions expanded use of polygraph. APA Monitor, 12-13.

Saal, F. E., Downey, R. G., & Lahey, M. A. (1980). Rating the ratings: Assessing the psychometric quality of rating data. Psychological Bulletin, 88, 413-428.

Saxe, L., Dougherty, D., & Cross, T. (1985). The validity of polygraph testing: Scientific analysis and public controversy. American Psychologist, 40, 355-366.

Sackett, P. R., & Decker, P. J. (1979). Detection of deception in the employment context: A review and critical analysis. Personnel Psychology, 32, 487-506.

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. Psychological Bulletin, 86, 420-428.

Spielberger, C. D., Gorsuch, R. L., & Lushene, R. E. (1970). Manual for the State Trait Anxiety Inventory, Palo Alto, CA: Consulting Psychologists Press.

Szucko, J. J., & Kleinmuntz, B. (1981). Statistical versus clinical lie detection. American Psychologist, 36, 488-496.

VanDercar, D. H., Greaner, V., Hibler, N. S., Spielberger, C. D., & Block, S. (1980). A description and analysis of the operation and validity of the Psychological Stress Evaluator. Journal of Forensic Science, 25, 174-188.

Author Notes

An earlier version of this article was presented at the Ninety Third annual convention of the American Psychological Association. We thank Patrick A. Knight for his helpful comments and thoughts on the article.

Requests for reprints should be sent to Ronald G. Downey, Department of Psychology, Bluemont Hall, Kansas State University, Manhattan, KS, 66506.

Table 1 - F-Values for the Chart by Rater by Ratee ANOVA for Each Question

	F-VALUES					
ITEM	CHART (C)	RATER (RR)	RATEE (RE)	C X RR	C X RE	RR X RE
1	8.62**	1.09	1.99*	0.14	1.27	1.17
2	0.01	8.23**	4.28**	0.65	2.02*	1.10
3	1.62	2.53	5.12**	0.71	6.94**	1.40
4	0.3	2.50	3.37**	1.41	1.83	0.92
5	8.56**	2.52**	3.61**	0.27	1.87	0.87
6	2.70	5.40**	3.40**	0.23	2.24*	0.65
7	0.16	1.18	5.34**	0.24	1.18	0.69
8	1.04	1.80	3.58**	0.33	2.17*	1.00
9	14.01**	3.53*	5.26*	0.37	1.66	1.01
10	8.90*	3.90**	2.10**	0.37	1.72	2.01
11	4.35	8.73*	4.27*	0.04	1.43	0.57
12	1.75	3.23	2.11	0.68	2.86**	0.87
13	0.01	2.39	1.44**	0.56	1.51**	1.37
14	5.88**	1.43**	4.72**	1.71	2.69	0.85
15	11.62**	4.59**	8.75**	0.48	1.45	0.53
16	15.32	7.04**	3.75**	0.87	1.15	0.77
17	3.31	8.65**	3.20**	0.13	0.99	0.89
18	3.15**	8.53**	5.30**	3.02*	1.71**	1.02
19	7.82**	14.30**	3.51**	1.20	2.60**	0.80
20	12.57	7.87	3.63**	1.52	2.95	1.07
21	0.03	1.75	3.35**	0.74	1.53	0.55
22	7.02*	4.51**	4.45**	0.88	1.27	0.85
23	0.04	16.07**	4.24**	0.44	2.07**	0.85
MULTIVARIATE	3.91	2.17	3.10	0.92	2.17	1.03

* $p < .05$: ** $p < .01$

Degrees of freedom: C=1, RR=3, RE=14, C X RR=3, C X RE=14, RR X RE=42, and ERROR=42.

Table 2 - Means, Standard Deviations, and ICCs by Item for Regular (R) and Telephone (T) Charts with the Correlations Between Conditions (R & T) - Ratings Averaged over Raters

ITEM #		MEAN	S.D.	ICC	r(R/T)	ITEM #		MEAN	S.D.	ICC	r(R/T)
					RELEVANT ITEMS (RQ)						
5	R	2.65[1]	0.80	0.35[2]		6	R	2.78	0.70	0.26[2]	
	T	2.15	0.69	0.23	0.26		T	2.50	0.71	0.34	-0.01
8	R	2.61	0.82	0.33[2]		9	R	2.93[1]	0.76	0.29[2]	3
	T	2.43	0.86	0.31	0.25		T	2.33	0.87	0.47	0.53
10	R	2.78[1]	0.66	0.24[4]		11	R	2.82[1]	0.81	0.38[4]	
	T	2.38	0.60	0.15	0.42		T	2.38	0.77	0.18	0.26
15	R	2.88[1]	0.79	0.27[2]	3	16	R	2.87[1]	0.78	0.29[4]	
	T	2.25	0.97	0.58	0.53		T	2.15	0.63	0.18	0.44
17	R	2.57	0.68	0.16[5]		18	R	3.00	0.75	0.32[2]	3
	T	2.23	0.71	0.25	0.48		T	2.70	0.98	0.43	0.54
19	R	2.85[1]	0.76	0.38[2]		20	R	2.95[1]	0.84	0.43[2]	
	T	2.36	0.80	0.30	0.04		T	2.40	0.73	0.30	0.14
22	R	2.70[1]	0.61	0.20[2]		23	R	2.62	0.60	0.15[5]	
	T	2.28	0.75	0.39	0.51		T	2.58	0.92	0.47	0.30
					KT & GC ITEMS (TG)						
1	R	2.20[1]	0.50	0.03[5]		2	R	2.15	0.67	0.37[2]	
	T	1.75	0.52	0.39	0.30		T	2.13	0.71	0.35	0.40
4	R	2.08	0.64	0.28[2]		12	R	2.45	0.47	0.09[5]	
	T	2.00	0.67	0.28	0.26		T	2.25	0.76	0.40	-0.25
13	R	2.25	0.56	0.13[5]		21	R	2.33	0.40	-0.08[5]	
	T	2.23	0.59	0.09	0.13		T	2.37	0.84	0.43	0.12
					OSI & KL ITEMS (LO)						
3	R	2.30	0.67	0.39[2]		7	R	2.53	0.77	0.45[2]	3
	T	2.45	1.01	0.64	0.02		T	2.46	0.66	0.20	0.52
14	R	2.65[1]	0.83	0.51[2]							
	T	2.25	0.82	0.31	0.20						

[1] Means were significantly different ($p<.05$) in the condition by rater by ratee ANOVA.

[2] Ratee effects (R & T) were significant ($p<.05$) - rater by ratee ANOVA.

[3] The correlation between R and T ratings was significant ($p<.05$).

[4] Same as footnote 2 but only the R ICC was significant.

[5] Same as footnote 2 but only the T ICC was significant.

Table 3 - Single Rater Reliabilities over the 'Relevant' Items with Means and Standard Deviations and Intercorrelations between Raters

RATER	MEAN	S.D.	RATER 1-R	2-R	3-R	4-R	1-T	2-T	3-T	4-T
1-REG	42.93	13.34	(0.93)[2]							
2-REG	41.53	13.07	0.21	(0.96)						
3-REG	39.27	4.91	0.62	0.46*	(0.60)					
4-REG	32.33	8.67	0.29	-0.03	0.33	(0.89)				
1-TEL[3]	34.07	14.46	0.71	0.33	0.57*	0.42	(0.96)			
2-TEL	39.00	14.34	0.25	-0.24	-0.11	0.22	0.28	(0.96)		
3-TEL	34.20	9.31	0.56*	0.28	0.44*	0.30	0.86**	0.26	(0.92)	
4-TEL	25.33	8.97	0.25	0.23	0.27	-0.31	0.40	0.13	0.50*	(0.90)

[1] REG = ratings of the regular charts.

[2] Values in () are ICCs for a single rater over the 'relevant' items

[3] TEL = ratings of the telephone charts.

* p < .05: ** p < .01